JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 9(2), JUNE, 2021
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com

Prediction of Heart Disease Using Machine Learning Algorithms

YUSUF Aminat Bolatito*, KIYEA Celinus
Department of Information and Communication Technology,
Usmanu Danfodiyo University, Sokoto, Nigeria

ABSTRACT
Disease diagnosis aids clinicians in making appropriate therapy suggestions for patients. Heart disease is one of the most frequent diseases today, and early detection of the condition is critical for many health care providers in order to protect their patients and save lives. In this sector, one of the solutions for detecting disease-related symptoms is to utilize machine learning. Medical datasets have made machine-learning algorithms more intuitive. Machine learning algorithms offer methods for identifying datasets of features and their interactions. Various machine-learning techniques are analyzed on heart disease datasets in attempts to accurately identify and or forecast cases of heart disease. In this study, seven parameter tuned machine learning algorithms were used to integrate and verify 14 variables of 920 patient data from Cleveland, Hungary, Switzerland, and Long Beach. K-Nearest Neighbor, Random Forest, Gaussian Naive Bayes, Logistic Regression, Support Vector Machine, Adaboost, and Gradient Boosting methods are among the hyperparameters tailored algorithms used. In order to compare the performance of the models, five performance metrics were used: accuracy, precision, sensitivity, F1 score, and Area under the Curve. The findings were compared to machine learning methods that used the same-pooled datasets. It was observed that using different classification algorithms for the classification of the combined dataset and tuning the parameters of these algorithms produced very promising results in terms of classification accuracy for the Adaboost, Support Vector Machine, Naive Bayes, and Gradient Boost classifiers, with classification accuracies of 85.98%, 85.87%, 84.56%, and 84.39%, respectively.

ARTICLE INFO
Article History
Received: December, 2020
Received in revised form: March, 2021
Accepted: June, 2021
Published online: July, 2021

KEYWORDS
Heart Diseases, datasets, Machine Learning Techniques, Parameters Tuning

INTRODUCTION

The heart is a sensitive organ that pumps and supplies blood to the various organs of an organism.

If the heart does not pump enough blood to the body's essential organs, symptoms such as shortness of breath, muscular weakness, and swollen feet might occur (Durairaj & Ramasamy, 2016). As a result, the heart must be in excellent working order to perform its functions. Nowadays, heart disease is recognized as a prevalent condition that causes heart failure, resulting in the death of many individuals and shortening peoples' life spans (Diwakar et al., 2021). The people who are the most impacted are those who are over 40 years old and above (Yang et al., 2015). According to the European Cardiology Society, 26 million individuals worldwide have been diagnosed with heart disease, with an additional 3.6 million diagnosed each year (Li et al., 2020). Approximately half of all patients diagnosed with heart disease die within 1-2 years (Haq et al., 2018) and treating heart disease consumes around 2% to 3% of the overall health-care expenditure (Haq et al., 2018; Katarya & Srinivas, 2020). In the same vein, the World Health Organization said in 2016 that heart disease accounts for almost 30% of all fatalities worldwide. That is, 17.90 million individuals die each year due to this disease throughout the world.

Corresponding author: Yusuf, A. B. ✉ aminbolly@gmail.com ✉ *Department of Information and Communication Technology, Usmanu Danfodio University, Sokoto.*

Coronary artery, vascular disease, heart rhythm abnormality, structural heart disease, and heart failure are some of the several forms of heart disorders (Katarya & Srinivas, 2020). Diabetes, high blood pressure, increased stress, cholesterol, obesity, age, and family history are the primary causes of heart disease (Tougui et al., 2020). Nonetheless, some lifestyle choices, such as smoking, eating an unhealthy diet, lacking in exercise and fitness, and so on, might raise the risk of heart disease (Das et al., 2009). Early and accurate identification of this fatal disease is critical for preventing further harm and preserving patients' lives.

For many years, experts have been trying to come up with an effective approach for assisting health care professionals in identifying some of the early symptoms of heart disease (Hassan et al., 2019). The existing traditional invasive-based methods have a number of flaws in terms of early disease detection (Allen et al., 2012). This is due to a number of factors, including the fact that analysis is based on the patient's medical history, physical examination report, and physician's interpretation of the symptoms. However, the findings of this diagnostic are insufficient in identifying heart disease patients. Furthermore, analysis is costly and computationally complex (Tsanas et al., 2011). Angiography is regarded as one of the most precise procedures for detecting heart abnormalities among the traditional techniques. Angiography, on the other hand, has certain disadvantages, including high cost, a variety of side effects, and a high level of technological expertise(Muhammad et al., 2020) is required. Researchers attempted to develop different non-invasive smart healthcare systems based on predictive machine learning techniques such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes (NB), and Decision Tree (DT) to overcome the issues in conventional invasive-based methods for the detection of heart disease (Muhammad et al., 2020). As a result, the death rate among individuals with heart disease has reduced (Methaila et al., 2014). The Cleveland heart disease dataset is widely used by researchers in the literature (Nazir et al., 2018; Samuel et al., 2017). In this context, (Shah et al., 2020) utilized the University of California at Irvine's (UCI) Cleveland heart disease dataset to predict heart disorders using SVM, DT, Linear Regression, and KNN algorithms, with accuracies of 83%, 79%, 78%, and 87%, respectively. Similarly, they used WEKA tool to classify heart disease using NB, DT, RF, and KNN, achieving accuracies of 88.157%, 90.789%, DT 80.263%, and RF 86.84%. Likewise, (Geetha et al., 2021) created a Heart Disease classification system using KNN, DT, SVM, and RF, with KNN achieving the highest accuracy of 87%.

There are still a lot of difficulties that might preclude reliable heart disease prediction, such as lack of medical databases and in-depth analysis (Detrano et al., 1989; Gennari et al., 1989). Large datasets are required for training and testing machine learning prediction models. The performance of a machine learning model may be assessed and compared to small datasets when huge datasets are explored. The model's performance improves when a balanced dataset is used for training and testing. As a result, data balance is critical for improving model performance. In this study, attention is paid to increasing medical databases and in-depth analysis of the datasets. Data preparation techniques are utilized for in-depth examination in order to increase the prediction power of machine learning models. Seven different machine learning classification methods were employed in this work, including KNN, RF, Guassian Naive Bayes (NB), Logisitic Regression (LR), SVM, Adaboost (AB), and Gradient Boosting (GB). Various preprocessing approaches were advocated, including imputation of missing feature value instances from the dataset, Standard Scalar, and dealing with the problem of categorical features. Also in the settings of suggested machine learning algorithms, the notion of grid search strategy was utilized. It is viewed as a distinct approach of obtaining the best parameter for any model in order for the classifier to accurately predict unlabeled data, i.e. testing data (Ramadhan et al., 2017).

REVIEW OF RELATED WORKS

Heart disease has been identified as one of the leading causes of mortality. Several study investigated the use of prediction models to foresee symptoms in heart disease patients based

Corresponding author: Yusuf, A. B. ✉ aminbolly@gmail.com ✉ *Department of Information and Communication Technology, Usmanu Danfodio University, Sokoto.*

on published health data. Recent advancements in machine learning tools and algorithms have aided in the development of methodologies and procedures for the detection of heart disease. In order to discover risks in the early stages of heart disease, data analysis is always required. The variance and bias of a dataset influence the quality of data analysis. Any algorithm's performance is influenced by the dataset's variance and bias(Sharma & Rizvi, 2017). The use of low variance and high biasness has various advantages, including the fact that it takes less time to train and test the algorithm. However, there are certain disadvantages to working with a limited dataset. As the dataset grows in size, the asymptotic errors appear to have a slight bias, according to the literature(Sharma & Rizvi, 2017). Buttressing this argument, the review focuses on the amount of data used in heart disease prediction utilizing machine learning algorithms, as well as the accuracy of the predictions.

Gokulnath & Shantharajah, (2019) created a heart disease prediction model. They used the Cleveland dataset's 303 samples with 14 input, and a genetic algorithm was used to perform feature selection. They obtained seven features as a result of this procedure, which were used to create heart prediction models using four machine-learning methods: SVM, Multilayer perception, J48, and KNN. The framework was also evaluated using the 10-fold cross validation method, and when the results were compared to models created using the original commonly used feature selection techniques. It was discovered that the SVM combined with the genetic algorithm had an accuracy of 88.34%, compared to 83.7% with the original dataset.

Almustafa, (2020) carried out research on the comparative analysis and classification of machine learning algorithms for the prediction of heart disease. A set of data from Cleveland, Hungary, Switzerland, and Long Beach was used in the study. KNN, NB, Decision tree J48, JRip, SVM, AB, Stochastic Gradient Decent, and Decision Table classifiers were used on the dataset to demonstrate the performance of the selected classification algorithms to best classify and predict cases with heart disease. The results demonstrate that utilizing multiple classification algorithms for heart disease classification yielded encouraging results in terms of accuracy, with accuracy classifications of 99.7073%, 98.0488%, and 97.2683% for KNN (N=1), Decision tree J48, and JRip, respectively. The accuracy of KNN (N=1) and Decision Table classifiers was improved when features were extracted using the classifiers subset Evaluator. Using a lesser number of features (4 out of 13) rather than all of the available attributes resulted in a more reliable feature selection technique for heart disease prediction.

In the same search for reliable heart disease prediction, (Spencer et al., 2020) employed four separate datasets that were then merged into one. Construct separate feature sets, they used Principal Component Analysis, Chi squared testing, ReliefF, and symmetrical uncertainty to analyze the dataset. Then, using various classification methods, a number of machine learning models were created. Their results were compared using a range of measures, including accuracy, recall, and precision. Chi-squared feature selection using BayesNet classifier increased accuracy by 85.0%, precision by 84.73%, and recall by 85.56% across the many experimental combinations utilized.

Similarly, (Dinesh et al., 2018) used data from four databases, including Hungarian, Cleveland, Virginia, and Long Beach, all of which were available through the UCI Repository. SVM, GB, RF, NB classifier, and LR were used to obtain better and more accurate prediction results. They utilize strategies to eliminate noisy data, removing missing records where applicable, and categorization of attributes for prediction and decision making at various stages during the data preparation step. On the basis of accuracy, sensitivity, and specificity analysis, the performance of each of these algorithms on the dataset was evaluated. SVM, GB, RF, NB, and LR produced total performance values of 79.7%, 84.2 percent, 80.8%, 84.2%, and 86.5%, respectively.

In their research, (Saxena & Sharma, 2016) created a model that encapsulates a DT and can accurately predict heart disease. Conducting the experiment, dataset from UCI in conjunction with attributes such as sex, age, chest pain, cholesterol, blood pressure, and many more were used. The dataset was split into two sections: training data and test data.

Corresponding author: Yusuf, A. B. ✉ *aminbolly@gmail.com* ✉ *Department of Information and Communication Technology, Usmanu Danfodio University, Sokoto.*

Javeed et al., (2019) addressed the issue of over-fitting, which resulted in improved accuracy on training data for predicting heart disease. This was accomplished by the creation of a model that included two algorithms: the Random Search Algorithm and the Random Forest Algorithm, both of which were used to forecast the model. The training and testing data, a superior result was attained.

A new technique for assessing the risk of heart disease has been presented as an alternative to existing methods(Weng et al., 2017). These researchers used RStudio to evaluate four basic classification algorithms on a dataset generated from the UK's Clinical Practice Research Datalink. RF, LR, Neural Network and GB machines were among the algorithms used, which were based on the American College of Cardiology (AHA) baseline model. The data was divided into two halves, 75% for training and 25% for testing. When the two models were compared, it was discovered that all machine learning methods outperformed the AHA model, with the neural network algorithm achieving the highest accuracy of 76.4%, followed by GB and LR with accuracies of 76.1% and 76.0%, respectively. The AHA model earned 74.5% for RF.

Khateeb & Usman, (2017) used the Cleveland heart-disease dataset to evaluate a number of classification algorithms, including NB, DT, KNN and Bagging. Instead of utilizing a feature selection algorithm to find the most statistically significant variables, they picked their features based on domain expertise. The accuracy of their NB and KNN approach model rose, but the accuracy of their DT and Bagging models declined. The model created by resampling the Machine learning KNN method using the original dataset has an accuracy of 79.2%.

MATERIALS AND METHOD

Data Gathering

The Heart Disease dataset used in this work is obtained from the UCI repository. This comprised of four different databases including: the Cleveland dataset created by the Cleveland Clinic Foundation, the Long-Beach-VA dataset created by the VA Medical Center in Long Beach, the Hungarian dataset created by the Hungarian Institute of Cardiology in Budapest, and the Switzerland dataset created by the University Hospitals of Zurich and Basel in Switzerland. The goal of merging the four datasets is to get a greater number of instances so that machine-learning techniques can generate more reliable prediction models. There are 14 features and 920 instances in the newly merged dataset. Table 1 shows the features and their descriptions.

Table 1: Various Features and their Description used for Classification

Features	Description	Data Type
age	Patient age in years	Numeric
sex	Patient gender	Nominal
cp	Chest Pain types of values: 1=Typical Agina 2=Atypical Agina 3=Non-anginal pain 4=Asymptomatic	Nominal
testbps	Value of Blood Pressure at resting mode	Numeric
chol	Serum Cholesterol	Numeric
fbs	Fasting Blood Sugar Levels	Nominal
restecg	Results of electrocardiogram while at rest	Nominal
thalach	The accomplishment of maximum rate of heart	Numeric
exang	Agina induced by exercise	Nominal

Corresponding author: Yusuf, A. B. ✉ *aminbolly@gmail.com* ✉ *Department of Information and Communication Technology, Usmanu Danfodio University, Sokoto.*

oldpeak	Exercise induced ST depression in comparison with state of	Numeric
slope	ST segment measurement	Nominal
ca	Fluoroscopy coloured major vessels	Numeric
thal	Status of heart through 3 values	Nominal
target	The predicted class	Nominal

Figure 1 shows the percentage of people with heart disease in the new data set, which is 55.33%. The dataset's class level distribution shows how much True/False positive data are there in the target variable. Out of them, 411 were positive (heart disease present), whereas 509 were negative (no heart disease present). The bar chart in Figure 1 depicts the number of heart disease and non-heart disease patients in the target class.

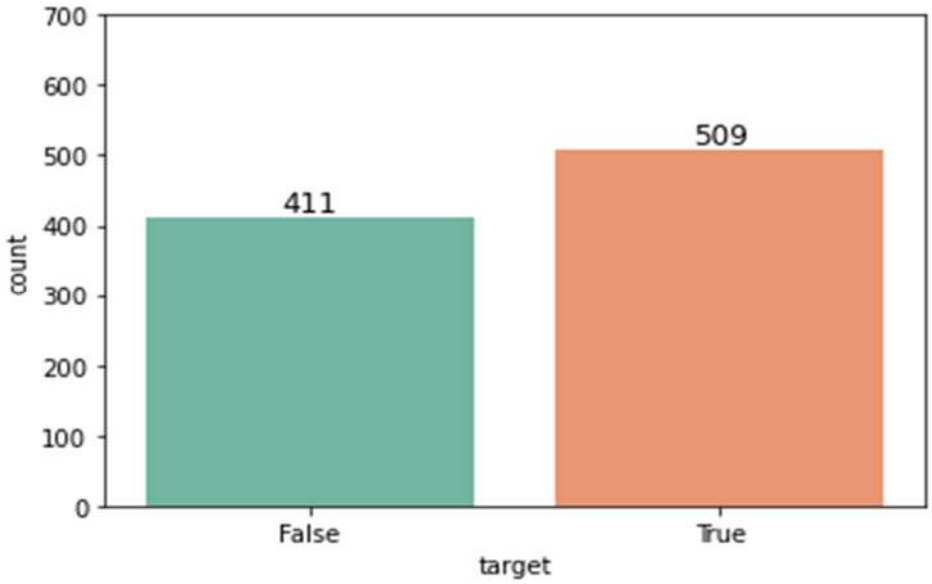

Figure 1: Dataset's class level showing presence and absence of heart disease

Figure 2 depicts the research architecture of the model for predicting the existence of heart disease. In the first phase of the procedure, the integrated heart disease datasets were preprocessed to manage missing values, and machine learning techniques were used to predict if a patient had heart disease or not. Finally, the algorithms were evaluated using performance metrics based on the confusion matrix. The workflow of recommended approaches for the implementation of Machine Learning Algorithms is depicted in Figure 2.

Corresponding author: Yusuf, A. B. ✉ *aminbolly@gmail.com* ✉ *Department of Information and Communication Technology, Usmanu Danfodio University, Sokoto.*

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 9(2), JUNE, 2021
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com

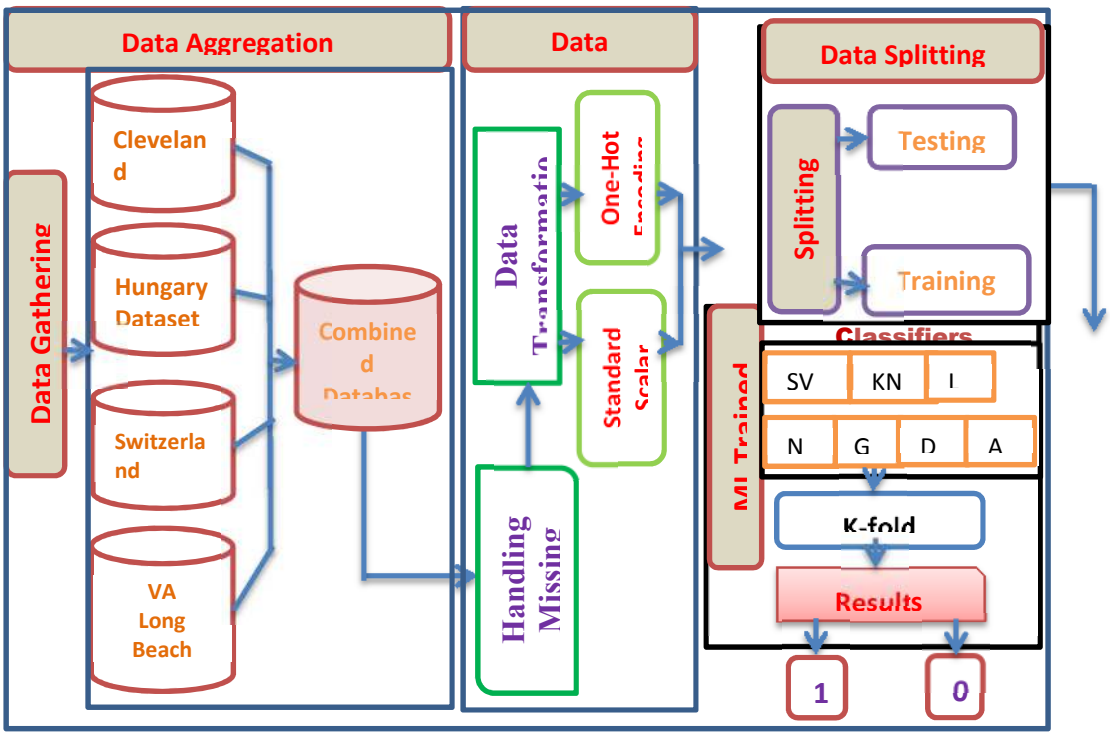

Figure 2: The research architecture

Data Preprocessing

The most important initial step in any predictive model is data preparation; it helps to put data into an intelligible format, which improves model's efficiency (Hamdaoui et al., 2020). Firstly, it was observation that the newly merged dataset has lot of missing values, especially in the ca, thal, and slope features. Because these missing values are categorical, calculation utilizing K-Nearest Neighbor Imputation was used to replace them with the nearest neighbor value.

Secondly, after investigation it was noted data set's projected class (target label) had five values: a value of 0 indicated no heart disease, while values between 1 and 4 indicated varying degrees of heart disease. The existence or absence of heart disease is the focus of this investigation, rather than the precise disease categorization. As a result, the class label were categorized as a binary value of 0 or 1, indicating whether the patients have heart disease or not.

Medical data is typically inadequate, lacking attribute values, and noisy due to outliers or unnecessary information (Zolbanin et al., 2015). Each parameter, particularly the numeric values, contributes to better disease prediction. Thirdly, it was discovered using distribution that the majority of the numeric columns were heavily skewed to the right. The problem was resolved by using log transform. Figure 3 illustrates this.

Corresponding author: Yusuf, A. B. ✉ aminbolly@gmail.com ✉ *Department of Information and Communication Technology, Usmanu Danfodio University, Sokoto.*

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 9(2), JUNE, 2021
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com

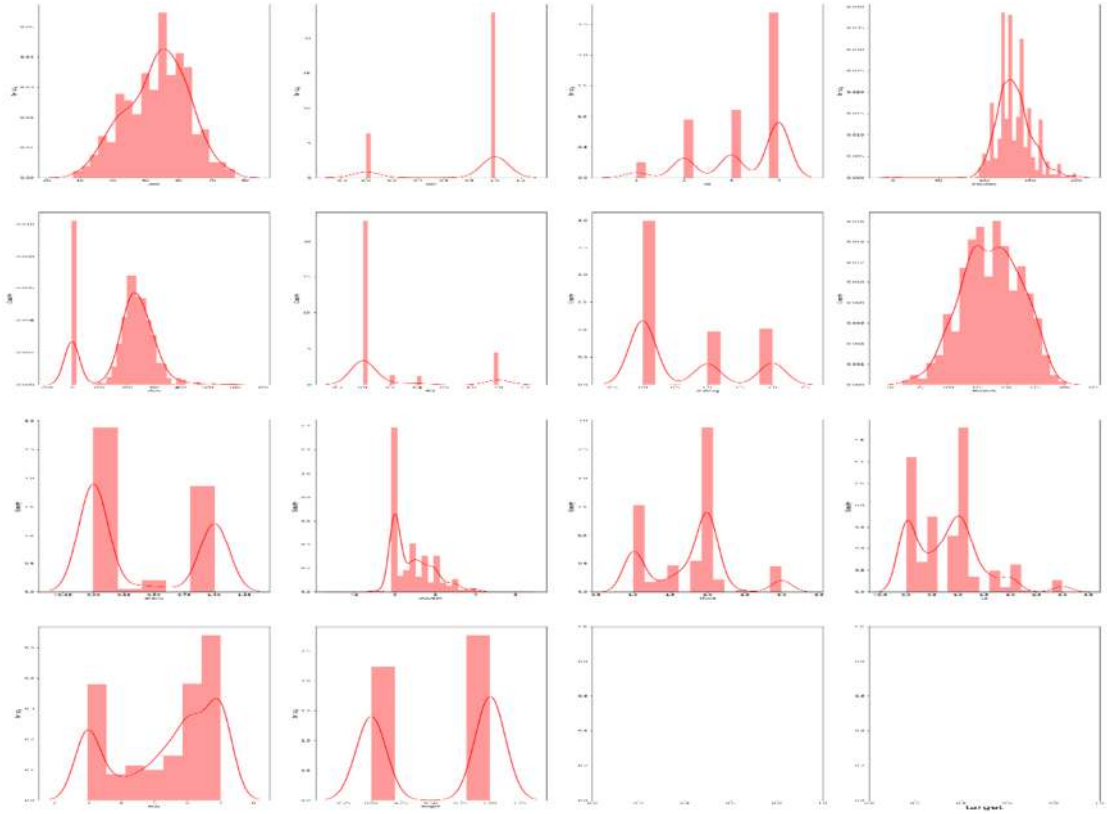

Figure 3: Distribution of the Nominal and Numeric Features

Lastly, there are a variety of features in the heart disease dataset, including numerical and nominal values. These factors exacerbate the difficulty of the computing process. As a result, features of numerical values in the range of zero to one are normalized using a normalization approach, as well as one hot encoding techniques used on nominal values. The goal is to reduce both the nominal and numerical complexity of the heart disease prediction algorithm. Data normalization may be done in a variety of ways. The well-known conventional normalizing approach is utilized in this investigation. Using the following equation, this technique plots a numerical value from the original dataset X in the interval [0, 1]:

$$\sigma = \frac{X - Min(X)}{Max(X) - Min(X)} \qquad 1$$

Where X is the original feature value and σ is the features' normalized value. If these factors are not correctly managed, they can have an influence on the quality, predictive power, and accuracy performance of the model.

Data Splitting

The goal of dividing the data is to avoid overfitting the model during model testing on the testing dataset. The dataset for this study was splitted into two parts: training data (80%) and test data (20%).

Machine learning classification algorithms.

Predictive modeling is a method for constructing a model that can make predictions. For the sake of producing those predictions, this model contains a machine learning technique that allows data-driven models to learn particular information from observed data in a training data

Corresponding author: Yusuf, A. B. ✉ aminbolly@gmail.com ✉ *Department of Information and Communication Technology, Usmanu Danfodio University, Sokoto.*

set (Alharthi, 2018). In this research, a specific target is utilized to forecast output for newly used instances, determine whether heart disease is present or not. As a result, the best way to train the data is to use a supervised learning classification method. Our suggested model is built using the following machine learning algorithms: NB, KNN, SVM, RF, LR, AB, and GB.

K-Nearest Neighbour classifier is a frequently used classification machine learning method for distinguishing between healthy and sick instances. It creates competitive results that will significantly enhance the current situation if it is intelligently coupled with past experience. The KNN rule classifies each unknown instance in the training set based on a majority of its KNN neighbors. Its performance is also heavily influenced by the distance measure that is used to determine the closest neighbors. To measure the difference between instances represented as vector inputs in the absence of previous knowledge, most K-NN classifiers employ basic Euclidean metrics. Other recommended distance measuring formulae include Xing distance computations, in addition to the traditional distance methods such as Minkowski and Chebyshev (Khorshid & Abdulazeez, 2021). A KNN categorizes an example to the class that appears the most frequently among its k close neighbors. K is a restriction for fine-tuning this work's categorization. In this work, the value $p = 2$ for manhattan_distance and n_neighbors value 5 were applied.

Support Vector Machines (SVMs) are a type of supervised machine learning technique used for classification and regression. Because of its capacity to cope with both continuous and categorical data, SVM is a popularly used method. It locates an appropriate hyper plane in N-dimensional space that unambiguously classifies the provided dataset with the least amount of data points between them (Almustafa, 2020). The goal of these methods is to solve issues without having to solve any intermediate problems. SVM extends the notion of structural risk reduction to include the ability to avoid the problem of over-fitting. For $y_i = 0, 1$ correspondingly, SVM was used for binary classification with two categories; absence and presence of heart disease. The goal of SVM is to divide the datasets into classes in order to find the most extreme peripheral hyper-plane. In this work the kernel "rbf" was used.

Random Forest classifier is a classifying technique that is capable of handling both regression and classification tasks. This classifier is an ensemble algorithm, which means it is made up of several decision tree algorithms. With the use of numerous decision trees and a technique called Bootstrap Aggregation, often known as bagging (Ghosh et al., 2021). During its training, it creates an entire forest of unrelated and random trees. Ensemble learners use a combination of algorithms to create an optimal predictive model that outperforms any of the individual models.

The Gradient Boosting technique aims at improving weak learning or poorly predictive ideas. The objective of GB is to construct a decision tree with a significantly higher connection by integrating many ideas with a weak predictive component and a smart algorithm. This idea is particularly useful with huge datasets if there are not a lot of concepts in common across the data pieces. One of the most common applications of these technologies is in search engine rankings. Search engine rankings must take in a huge number of possible queries that may or may not be linked to one another and divide them into a small number of rankable terms.

Logistic regression is a real-valued input vector discriminative classification approach. Features or predictors are the measurements of the input vector to be classified. Multiclass classification can also be used with LR. In LR, the probability P of a dichotomous event may be deduced from the Bernoulli trial and linked to the event under investigation (Hajmeer & Basheer, 2003; Schumacher et al., 1996; Vach et al., 1996). The solver and the maximum number of iterations were the two LR parameters employed in this investigation. The algorithm used in the optimization problem is known as the solver. The number of iterations required for the solvers to converge is specified between 100 and 500.

Gaussian Naive Bayes networks are made up of network-like structures with conditional probabilities to go along with them. The NB is a directed acyclic graph with nodes that represent provincial variables and edges between nodes that represent variable dependences. Two NB parameters priors and Var_smoothing were

Corresponding author: Yusuf, A. B. ✉ aminbolly@gmail.com ✉ *Department of Information and Communication Technology, Usmanu Danfodio University, Sokoto.*

chosen in this investigation. Priors represent the prior probabilities of the classes. If this parameter is specified while fitting the data, the prior probability will not be justified by the data.

Adaboost, or Adaptive Boosting, is a binary classification boosting method that combines a number of weak classifiers to create a more robust classifier. Based on 1000 samples, this method calculates the anticipated accuracy. The instances in the training dataset are given a starting weight (Karim et al., 2019).

Classifier validation

In machine learning, validation of the prediction model is a crucial stage. The findings of the above-mentioned classification models are validated using the k-fold cross-validation approach. In k-fold cross validation (CV), the entire dataset is separated into k equal portions of k. At each iteration, every (k-1) portion is used for training and the rest is used for testing. This procedure is repeated for k-iterations. Different researchers have employed different values of k for k-fold cross validation. The value k = 5 is applied in this study since it gives decent results for the experimental work. In fivefold CV, 95% of the data is used to train the model, while the remaining 5% is used to test the model at each iteration. Finally, the average of the findings from each stage is calculated to yield the results.

Performance evaluation metrics

Using the confusion matrix, different metrics such as sensitivity, accuracy, and precision were generated to assess the validity of the prediction models as shown in Table 2. The percentage of correctly identified actual positives is measured by sensitivity and the likelihood of a total number of correct predictions is defined by the accuracy. Whereas, the positive predictive value, is the precision (Stehman, 1997). The following formulae are used to describe these measurements numerically as seen in equations 2, 3, 4 and 5. Where TP, TN, FP, and FN stand for True Positive (number of positive data correctly labeled by the classifier), True Negative (number of negative data correctly labeled by the classifier), False Positive (number of negative data incorrectly labeled as positive), and False Negative (number of positive data incorrectly labeled as negative) respectively

.

Table 2: Confusion Matrix

		Actual Values	
		Positive	**Negative**
Predicted Values	**Positive**	TP	FP
	Negative	FN	TN

$$Accuracy = \frac{TP + TN}{(TP + FP + TN + FN)} \qquad equation\ 2$$

$$Precision = \frac{TP}{(TP + FP)} \qquad eqation\ 3$$

$$Recall\ or\ Sensitivity = \frac{TP}{(TP + FN)} \qquad equation\ 4$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad equation\ 5$$

Another classification metric is the ROC (Receiver Operating Curve) (Metz, 1978), which is a curve based on the FP ratio (x-axis) and the TP ratio (y-axis). For the best classifier, the Area under the ROC must be close to 1.

RESULT AND DISCUSSION

Using 13 variables described in the methodology, the performance of the seven machine learning algorithms for the prediction of heart disease was evaluated. A total of 920 cases

Corresponding author: Yusuf, A. B. ✉ aminbolly@gmail.com ✉ *Department of Information and Communication Technology, Usmanu Danfodio University, Sokoto.*

were examined, with 411 showing symptoms of heart disease and 509 showing no symptoms of heart disease. KNN, RF, NB, LR, SVM, AB, and GB were all used along with full feature vector in evaluating the performance metric of accuracy, precision, sensitivity, F1 score, and Area Under the Curve (AUC). During cross validation, data instances were partitioned into fivefolds, with each fold being used for testing and the remaining folds being used for training.

Table 3: Results presentation

	KNN	RF	NB	LR	SVM	AB	GB
Accuracy	88.04	84.24	84.56	82.07	85.87	85.98	84.39
Precision	89.91	88.46	90	87.25	87.39	87.58	86.87
Sensitivity	89.91	84.4	82.57	81.65	88.99	89.12	78.9
F1 Score	88	84	85	82	86	86	84
AUC	94	99	89	91	95	96	98

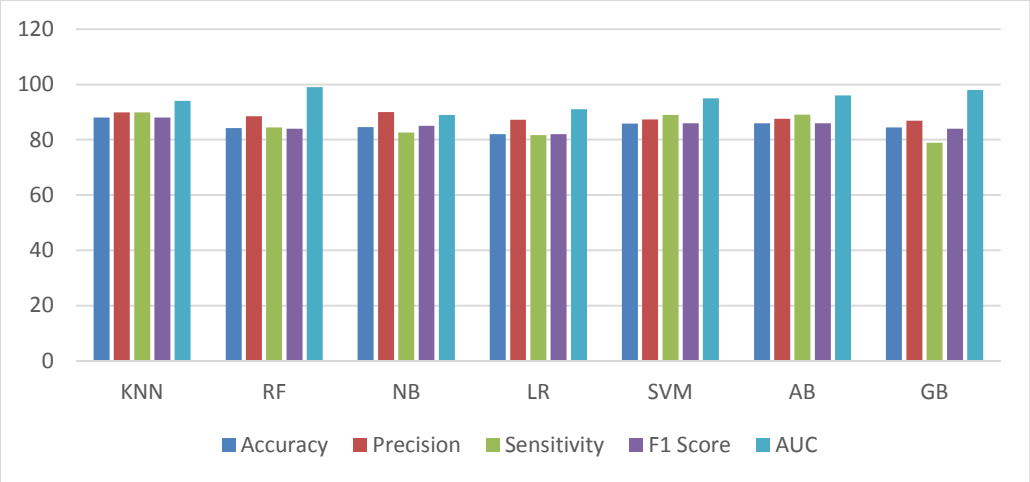

Figure 4: Algorithmic Comparison of different metrics

Table 3 details several classification recital metrics, whereas Figure. 4 depicts five classification performance metrics: accuracy, precision, sensitivity, F1 score, and AUC. Figure 4 shows that KNN beat all other machine learning algorithms, achieving the greatest classification accuracy of 88.04%, while AB (85.98%) earned the second best classification accuracy. Furthermore, these two techniques have a sensitivity of 89.91 and 89.12%, respectively. The classifier NB obtains the maximum precision of 90%, indicating that it is most suited for identifying individuals with heart disease (presence class). The graph clearly shows that RF has a 99% AUC, which outperforms all other methods. LR has the lowest F1 score (82%) while KNN has the greatest F1 score (88%).

Comparative Study

Our findings were compared to other models that have been developed on similar datasets. In earlier investigations, the greatest accuracies reported were 99.71% (Almustafa, 2020) in KNN methods, compared to our finding of 88.04%.

Corresponding author: Yusuf, A. B. ✉ *aminbolly@gmail.com* ✉ *Department of Information and Communication Technology, Usmanu Danfodio University, Sokoto.*

Table 4: Comparison of the Accuracy with related state of Art

Authors	Dataset	Methods	Accuracy
Proposed model	UCI Cleveland, Hungary, Switzerland, and Long Beach Dataset	KNN, RF, NB, LR, SVM, AB, GB	88.04%, 84.24%, 84.56%, 82.07%, 85.87%, 85.98%, 84.39%
(Spencer et al., 2020)	UCI Cleveland, Hungary, Switzerland, and Long Beach Dataset	RF, LR	83.91%, 83.91%
(Almustafa, 2020)	UCI Cleveland, Hungary, Switzerland, and Long Beach Dataset	KNN, NB, SVM, AB	99.71% , 83.12%, 84.20%, 84.29%
(Dinesh et al., 2018)	UCI Cleveland, Hungary, Switzerland, and Long Beach Dataset	RF, NB, LR, SVM, GB	80.90%, 84.27%, 86.52%, 79.78%, 84.27%

The similarity between our suggested model and other recent research efforts is shown in Table 4. The acquired results for the four datasets demonstrate a percentage improvement with some techniques with a corresponding drop in performance of a given algorithm, such as LR, which had an accuracy of 85.52%t (Dinesh et al., 2018). The suggested work recorded the best achievement with AB, which has an accuracy of 85.98%. In comparison with (Spencer et al., 2020) an accuracy of 84.20% and (Dinesh et al., 2018), 80.90%, the RF approach, demonstrates a consistent improvement in an accuracy of 84.24%.

CONCLUSION

Heart disease prediction has the potential to save lives and has a substantial influence on therapy. This study presents a process for predicting the presence or absence of heart disease based on machine learning approaches. This study compared seven traditional machine learning algorithms for predicting heart disease based on data preprocessing and various fine-tuning factors. These approaches were tested and verified using five-fold cross validation and a number of performance metrics. Furthermore, this effort may be extended for the prediction of heart disease by collecting vast amounts of data from multiple clinics, and employing feature selection techniques to reduce some of variables, resulting in a more comprehensive model. The recitation can be used to make operational predictions about heart disease.

REFERENCES

Alharthi, H. (2018). Healthcare predictive analytics: An overview with a focus on Saudi Arabia. *Journal of Infection and Public Health*, *11*(6), 749–756. https://doi.org/10.1016/j.jiph.2018.02.005

Allen, A., Larry, Kathleen, L. G., Daniel, D. M., & Felker, G. M. (2012). *Decision Making in Advanced Heart Failure*. 25. https://doi.org/DOI: 10.1161/CIR.0b013e31824f2173

Almustafa, K. M. (2020). Prediction of heart disease and classifiers' sensitivity analysis. *BMC Bioinformatics*, *21*(1), 278. https://doi.org/10.1186/s12859-020-03626-y

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease. *American Journal of Cardiology*, *64*, 7.

Corresponding author: Yusuf, A. B. ✉ *aminbolly@gmail.com* ✉ *Department of Information and Communication Technology, Usmanu Danfodio University, Sokoto.*

Dinesh, K., G., Santhosh, K. D., Arumugaraj, K., & Mareeswari, V. (2018). *Prediction of Cardiovascular Disease Using Machine Learning Algorithms*. 7.

Diwakar, M., Tripathi, A., Joshi, K., Memoria, M., Singh, P., & kumar, N. (2021). Latest trends on heart disease prediction using machine learning and image fusion. *Materials Today: Proceedings*, *37*, 3213–3218. https://doi.org/10.1016/j.matpr.2020.09.078

Durairaj, M., & Ramasamy, N. (2016). A Comparison of the Perceptive Approaches for Preprocessing the Data Set for Predicting Fertility Success Rate. *Int. J. Control Theory Appl.*, *9*, 255–260.

Geetha, K., Anitha, V., Elhoseny, M., Kathiresan, S., Shamsolmoali, P., & Selim, M. M. (2021). An evolutionary lion optimization algorithm-based image compression technique for biomedical applications. *Expert Systems*, *38*(1). https://doi.org/10.1111/exsy.12508

Gennari, J. H., Langley, P., & Fisher, D. (1989). Models of Incremental Concept Formation. *CONCEPT FORMATION*, 51.

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Mehedi, F. M. J. S., Ignatious, E., Shultana, S., Beeravolu, A. R., & Boer, D. F. (2021). Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques. *IEEE Access*, *9*, 19304–19326. https://doi.org/10.1109/ACCESS.2021.3053759

Gokulnath, C. B., & Shantharajah, S. P. (2019). An optimized feature selection based on genetic approach and support vector machine for heart disease. *Cluster Computing*, *22*(S6), 14777–14787. https://doi.org/10.1007/s10586-018-2416-4

Hajmeer, M., & Basheer, I. (2003). Comparison of logistic regression and neural network-based classifiers for bacterial growth. *Food Microbiology*, *20*(1), 43–55. https://doi.org/10.1016/S0740-0020(02)00104-1

Hamdaoui, H. E., Boujraf, S., Chaoui, N. E. H., & Maaroufi, M. (2020). A Clinical support system for Prediction of Heart Disease using Machine Learning Techniques. *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 1–5. https://doi.org/10.1109/ATSIP49331.2020.9231760

Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. *Mobile Information Systems*, *2018*, 1–21. https://doi.org/10.1155/2018/3860146

Hassan, Khourdifi, Y., Bahaj, M., & Hassan 1st University. (2019). Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization. *International Journal of Intelligent Engineering and Systems*, *12*(1), 242–252. https://doi.org/10.22266/ijies2019.0228.24

Javeed, A., Zhou, S., Yongjian, L., Qasim, I., Noor, A., & Nour, R. (2019). An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection. *IEEE Access*, *7*, 180235–180243. https://doi.org/10.1109/ACCESS.2019.2952107

Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K., & Alazab, M. (2019). A Comprehensive Survey for Intelligent Spam Email Detection. *IEEE Access*, *7*, 168261–168295. https://doi.org/10.1109/ACCESS.2019.2954791

Katarya, R., & Srinivas, P. (2020). Predicting Heart Disease at Early Stages using Machine Learning: A Survey. *2020 International Conference on Electronics and Sustainable Communication*

Corresponding author: Yusuf, A. B. ✉ aminbolly@gmail.com ✉ *Department of Information and Communication Technology, Usmanu Danfodio University, Sokoto.*

Systems (ICESC), 302–305. https://doi.org/10.1109/ICESC48915.2020.9155586

Khateeb, N., & Usman, M. (2017). Efficient Heart Disease Prediction System using K-Nearest Neighbor Classification Technique. *Proceedings of the International Conference on Big Data and Internet of Thing - BDIOT2017*, 21–26. https://doi.org/10.1145/3175684.3175703

Khorshid, S. F., & Abdulazeez, A. M. (2021). *BREAST CANCER DIAGNOSIS BASED ON K-NEAREST NEIGHBORS: A REVIEW*. 26.

Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. *IEEE Access*, *8*, 107562–107582. https://doi.org/10.1109/ACCESS.2020.3001149

Methaila, A., Kansal, P., Arya, H., & Kumar, P. (2014). Early Heart Disease Prediction Using Data Mining Techniques. *Computer Science & Information Technology (CS & IT)*, 53–59. https://doi.org/10.5121/csit.2014.4807

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*(4), 283–298. https://doi.org/10.1016/S0001-2998(78)80014-2

Muhammad, Y., Tahir, M., Hayat, M., & Chong, K. T. (2020). Early and accurate detection and diagnosis of heart disease using intelligent computational model. *Scientific Reports*, *10*(1), 19747. https://doi.org/10.1038/s41598-020-76635-9

Nazir, S., Shahzad, S., Mahfooz, S., & Nazir, M. (2018). *Fuzzy Logic based Decision Support System for Component Security Evaluation*. *15*(2), 8.

Ramadhan, M. M., Sitanggang, I. S., Nasution, F. R., & Ghifari, A. (2017). *Parameter Tuning in Random Forest Based on Grid Search Method for Gender Classification Based on Voice Frequency*. 6.

Samuel, O. W., Asogbon, G. M., Sangaiah, A. K., Fang, P., & Li, G. (2017). An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. *Expert Systems with Applications*, *68*, 163–172. https://doi.org/10.1016/j.eswa.2016.10.020

Saxena, K., & Sharma, R. (2016). Efficient Heart Disease Prediction System. *Procedia Computer Science*, 8.

Schumacher, M., Roßner, R., & Vach, W. (1996). Neural networks and logistic regression: Part I. *Computational Statistics & Data Analysis*, *21*(6), 661–682. https://doi.org/10.1016/0167-9473(95)00032-1

Shah, D., Patel, S., & Bharti, S. K. (2020). Heart Disease Prediction using Machine Learning Techniques. *SN Computer Science*, *1*(6), 345. https://doi.org/10.1007/s42979-020-00365-y

Sharma, H., & Rizvi, M. A. (2017). Prediction of Heart Disease using Machine Learning Algorithms: A Survey. *International Journal on Recent and Innovation Trends in Computing and Communication*, *5*(8), 6.

Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. *DIGITAL HEALTH*, *6*, 205520762091477. https://doi.org/10.1177/2055207620914777

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62*(1), 77–89. https://doi.org/10.1016/S0034-4257(97)00083-7

Tougui, I., Jilbab, A., & El Mhamdi, J. (2020). Heart disease classification using data mining tools and machine learning techniques. *Health and Technology*, *10*(5), 1137–1144.

Corresponding author: Yusuf, A. B. ✉ *aminbolly@gmail.com* ✉ *Department of Information and Communication Technology, Usmanu Danfodio University, Sokoto.*

https://doi.org/10.1007/s12553-020-00438-1

Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2011). Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. *Journal of the Royal Society Interface*, *8*(59), 842–855. https://doi.org/10.1098/rsif.2010.0456

Vach, W., Roßner, R., & Schumacher, M. (1996). Neural networks and logistic regression: Part II. *Computational Statistics & Data Analysis*, *21*(6), 683–701. https://doi.org/10.1016/0167-9473(95)00033-X

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS ONE*, *12*(4), e0174944. https://doi.org/10.1371/journal.pone.0174944

Yang, E. Y., Chambless, L., Sharrett, A. R., Virani, S. S., Liu, X., Tang, Z., Boerwinkle, E., Ballantyne, C. M., & Nambi, V. (2015). Carotid Arterial Wall Characteristics Are Associated With Incident Ischemic Stroke But Not Coronary Heart Disease in the Atherosclerosis Risk in Communities (ARIC) Study. *Stroke*, *43*(1), 103–108. https://doi.org/10.1161/STROKEAHA.111.626200

Zolbanin, H. M., Delen, D., & Hassan Zadeh, A. (2015). Predicting overall survivability in comorbidity of cancers: A data mining approach. *Decision Support Systems*, *74*, 150–161. https://doi.org/10.1016/j.dss.2015.04.003

Corresponding author: Yusuf, A. B. ✉ *aminbolly@gmail.com* ✉ *Department of Information and Communication Technology, Usmanu Danfodio University, Sokoto.*